Enhancing confidence in the detection of gravitational waves from compact binaries using signal coherence

Maximiliano Isi,[1,*] Rory Smith,[1,2,3,†] Salvatore Vitale,[4,‡] T. J. Massinger,[1] Jonah Kanner,[1] and Avi Vajpeyi[5,1]

[1]*LIGO, California Institute of Technology, Pasadena, California 91125, USA*
[2]*Monash Centre for Astrophysics, School of Physics and Astronomy, Monash University, Victoria 3800, Australia*
[3]*OzGrav: The ARC Centre of Excellence for Gravitational-Wave Discovery, Monash University, Victoria 3800, Australia*
[4]*LIGO, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[5]*Physics Department, The College of Wooster, Wooster, Ohio 44691, USA*

(Received 27 March 2018; published 27 August 2018)

We show that gravitational-wave signals from compact binary mergers may be better distinguished from instrumental noise transients by using Bayesian models that look for signal coherence across a detector network. This can be achieved even when the signal power is below the usual threshold for detection. This method could reject the vast majority of noise transients, and therefore increase sensitivity to weak gravitational waves. We demonstrate this using simulated signals, as well as data for GW150914 and LVT151012. Finally, we explore ways of incorporating our method into existing Advanced LIGO and Virgo searches to make them significantly more powerful.

DOI: 10.1103/PhysRevD.98.042007

I. INTRODUCTION

A pair of neutron stars or black holes merges somewhere in the observable Universe roughly every 15–200s, releasing large amounts of energy in the form of gravitational waves (GWs) [1–7]. One of the limiting factors in detecting such GWs with existing detectors, like Advanced LIGO (aLIGO) and Virgo [8,9], is data contamination by instrumental noise transients (glitches) that may mimic astrophysical signals [10]. Glitches can lower the inferred statistical significance of GW signals, making their detection more difficult. In this paper, we show how signal coherence may be used to address this problem by significantly improving our ability to distinguish genuine GW signals from glitches using Bayesian model comparison.

In particular, we demonstrate that Bayesian models—as proposed in [11]—may successfully distinguish real GWs from glitches by using the fact that the former must be *coherent* across detectors, while the latter will generally not be. Here, coherence means that a real GW must produce strain signals in different instruments that (i) are coincident in time (up to a time-of-flight delay), (ii) are well described by a compact-binary-coalescence (CBC) waveform, and (iii) share a phase evolution consistent with a single astrophysical source. In contrast, glitches should not be expected to fully satisfy these criteria. Making full use of this information—the expected coherence of signals and incoherence of glitches—may allow us to detect weaker signals than is currently possible.

From a subset of glitches and detection candidates (triggers) from aLIGO's first observation run (O1), we find that (a) the majority of glitches are markedly more incoherent than coherent across detectors, irrespective of their loudness or the detection significance assigned by one of the main detection pipelines; (b) simulated signals can be identified by their coherence, as long as they are distinguishable from Gaussian noise in at least two detectors; and finally, (c) the "gold-plated" detection GW150914 (detection significance $> 5.1\sigma$) [1] and the "silver-plated" candidate LVT151012 (detection significance $\sim 2.1\sigma$) [3] are both decidedly more coherent than incoherent. This study of real data thus implies that the Bayesian comparison of coherent and incoherent signal models has the potential to significantly improve the sensitivity of CBC searches, even with currently available computational resources.

II. SEARCHES

Templated searches for transient gravitational waves work by constructing a ranking statistic based on matched filtering [12–17]. In principle, to make a rigorous statement about the statistical significance of a pair of time-coincident triggers, it is necessary to know the probability that a given event was produced by instrumental noise, rather than an actual GW. This likelihood may be estimated empirically from the value of the ranking statistic for a large

*misi@ligo.caltech.edu
†rory.smith@ligo.org
‡salvatore.vitale@ligo.mit.edu

representative set of triggers known with certainty to be spurious. Such a set of signal-free triggers is denoted "background," in contrast to the "foreground" of candidates that may contain a signal.

Because detectors cannot be physically shielded from gravitational waves, *ad hoc* data analysis techniques must be used to estimate the background. One such strategy is to construct time slides by applying relative time offsets (longer than the light-travel time between sites) between the data of different detectors [16,17]. Detection significance can then be inferred, in a frequentist way, by comparing the value of the ranking statistic for a time-coincident foreground trigger to that of time-slid background triggers. The rate at which background triggers are produced with a given value of the ranking statistic is usually referred to as the "false-alarm rate" (FAR).

Efficient signal detection requires a ranking statistic that extracts the most information from the data, in order to discriminate between noise and weak astrophysical signals. However, existing CBC searches are *not* optimal in this sense: they do not incorporate knowledge of *all* features that may distinguish GWs from noise. Moving towards an optimal statistic is a great challenge, but one large step is to demand that foreground triggers in two or more detectors should be *better* described as coherent gravitational-wave signals, rather than incoherent glitches. Importantly, it is not enough to provide some measure of coherence: one must also prove that an incoherent model is not *more* successful at describing the data.

III. COHERENCE VS INCOHERENCE

To achieve this, we introduce the Bayesian coherence ratio (BCR): the odds between the hypothesis that the data comprise a coherent CBC signal in Gaussian noise ($\mathcal{H}_{\rm S}$), and the hypothesis that they instead comprise incoherent instrumental features ($\mathcal{H}_{\rm I}$)—meaning each detector has *either* a glitch in Gaussian noise ($\mathcal{H}_{\rm G}$) *or* pure Gaussian noise ($\mathcal{H}_{\rm N}$). For a network of D detectors,

$$\mathrm{BCR} \equiv \frac{\alpha Z^S}{\prod_{i=1}^{D} [\beta Z_i^G + (1-\beta) Z_i^N]}, \tag{1}$$

where Z^S is the evidence for $\mathcal{H}_{\rm S}$, and $Z_i^{\rm G}$ and $Z_i^{\rm N}$ are, respectively, the evidences for $\mathcal{H}_{{\rm G}i}$ and $\mathcal{H}_{{\rm N}i}$ in the ith detector. The arbitrary weights α and β parametrize our prior belief in each model: $\alpha = P(\mathcal{H}_{\rm S})/P(\mathcal{H}_{\rm I})$ and $\beta = P(\mathcal{H}_{{\rm G}i}|\mathcal{H}_{\rm I}) = 1 - P(\mathcal{H}_{{\rm N}i}|\mathcal{H}_{\rm I})$ for all i [see, e.g., Eq. (59) in [18]]. These priors will be chosen to minimize overlap between the signal and noise trigger populations; their importance is studied in detail in Appendix.

Evidences (marginalized likelihoods) are the conditional probability (P) of observing some data ($\mathbf{d}_i$, for detector i) given some hypothesis ($\mathcal{H}$). For the coherent-signal hypothesis, this is

$$\begin{aligned} Z^S &\equiv P(\{\mathbf{d}_i\}_{i=1}^D|\mathcal{H}_{\rm S}) \\ &= \int p(\vec{\theta}|\mathcal{H}_{\rm S}) p(\{\mathbf{d}_i\}_{i=1}^D|\vec{\theta},\mathcal{H}_{\rm S}) \mathrm{d}\vec{\theta}. \end{aligned} \tag{2}$$

The vector $\vec{\theta}$ represents a point in the space of parameters that describe the CBC signal, such as the component masses and spins; the terms in the integrand are the prior, $p(\vec{\theta}|\mathcal{H}_{\rm S})$, and the multidetector likelihood, $p(\{\mathbf{d}_i\}_{i=1}^D|\vec{\theta},\mathcal{H}_{\rm S}) = \prod_{i=1}^D p(\mathbf{d}_i|\vec{\theta},\mathcal{H}_{\rm S})$. The specific functional form of the single-detector likelihood, $p(\mathbf{d}_i|\vec{\theta})$, is derived from the statistical properties of the noise (e.g., a normal distribution for a Gaussian process). The integral is performed numerically using algorithms like nested sampling [19,20]. In our case, the data $\mathbf{d}_i$ are the calibrated Fourier-domain output of each detector, but could generally be any sufficient statistic produced from it.

Because of their inherently unpredictable nature, it is impossible to produce a template that *a priori* captures all features of a glitch. Therefore, we define a surrogate glitch hypothesis by the presence of simultaneous, but incoherent, CBC-like signals in different detectors. Thus, for the ith detector, the glitch evidence is

$$\begin{aligned} Z_i^G &\equiv P(\mathbf{d}_i|\mathcal{H}_{\rm G}) \\ &= \int p(\vec{\theta}_i|\mathcal{H}_{\rm G}) p(\mathbf{d}_i|\vec{\theta}_i,\mathcal{H}_{\rm G}) \mathrm{d}\vec{\theta}_i, \end{aligned} \tag{3}$$

where now we allow for a different set of signal parameters $\vec{\theta}_i$ at each detector.[1] We will set $p(\vec{\theta}_i|\mathcal{H}_{\rm G}) = p(\vec{\theta}_i|\mathcal{H}_{\rm S})$ and $p(\mathbf{d}_i|\vec{\theta}_i,\mathcal{H}_{\rm G}) = p(\mathbf{d}_i|\vec{\theta}_i,\mathcal{H}_{\rm S})$, but this may be relaxed to better capture specific glitch features, if necessary. The surrogate $\mathcal{H}_{\rm G}$ model captures the portion of glitches that lie within the manifold of CBC signals and, in a sense, corresponds to the worst possible glitch—one that looks exactly like coincident CBC signals. Variations of this strategy have been used before in the analysis of compact binary coalescences [11], minimally modeled transients [24–26], and continuous waves [27–29]. Other searches also make use of likelihood ratios in the detection process, but they do not rely on signal coherence (e.g., [13,14]).

Finally, because we assume a perfect measurement of the detector noise power-spectral density (PSD), the Gaussian-noise evidence is just the usual null likelihood. For our Fourier-domain data, this is just

$$Z_i^N \equiv P(\mathbf{d}_i|\mathcal{H}_{\rm N}) = \mathcal{N}(\mathbf{d}_i), \tag{4}$$

[1]Note that $\mathcal{H}_{\rm S}$ and $\mathcal{H}_{\rm I}$ are disjoint even if we do not explicitly exclude points from the parameter space satisfying $\vec{\theta}_i = \vec{\theta}_j$ for all $i \neq j$, because this condition defines a subspace that offers infinitesimal support to the prior in $\mathcal{H}_{\rm I}$ (see [18,21], or more general discussions in Ch. 4 in [22] or Ch. 28 in [23]).

where $\mathcal{N}(\mathbf{d}_i)$ is a multidimensional normal distribution with zero mean and variance derived from the noise PSD [20]. In principle, this could be easily generalized to marginalize over poorly known PSD parameters if needed.

IV. ANALYSIS

During O1, the two aLIGO detectors operated from September 12, 2015 to January 19, 2016. Ideally, we would like to compute the BCR for all triggers produced during this period to show that it can efficiently discriminate between glitches and CBC signals. However, computational limitations prevent this.[2] Instead, we pick a subset of 983 multidetector background binary-black-hole triggers identified by PyCBC, one of the staple search pipelines [15–17,30]. We pick the background triggers by sampling from the full trigger set uniformly in the log of the inverse-FAR (IFAR $\equiv 1/\mathrm{FAR}$) for IFARs in $[5\times10^{-5}, 10^{6}]$ yr, which is the total range reported by the pipeline. This sampling allows us to analyze common (low IFAR) and rare (high IFAR) background events.

To compute the evidences making up the BCR, Eq. (1), we run the nested-sampling algorithm implemented in the LALInference library on 4s-long data segments containing each trigger [20,31]. Given the large number of triggers involved, this would not be feasible without the reduction in the computational cost of Bayesian inference provided by reduced order quadrature (ROQ) methods (see, e.g., [32]). Using this technique makes no measurable difference for the values of the computed evidence.[3]

Templates are produced using IMRPhenomP, a standard waveform family [32–35]. We restrict the priors on the masses such that we only consider signals that are less than 4s in duration, resulting in a chirp-mass range of $12.3M_\odot \leq \mathcal{M} \leq 44.7M_\odot$. We further restrict the mass ratio to lie within $1 \leq q \leq 8$. The dimensionless spin magnitudes are taken to be within [0,0.89], and we consider all spin angles. The prior on luminosity distance assigns probability uniformly in volume, with an upper cutoff of 5 Gpc. These priors, as well as the priors for all other parameters, follow the default for standard LALInference analyses with ROQ [20,32]. The PSD used for matched filtering is calculated using the BayesWave algorithm [36,37].

The search that originally produced our set of triggers considered a wider range of masses and spins than we do in the BCR computation for the purpose of this demonstration. To accommodate this, we prescreened the background to only allow triggers with masses within our priors. It would be straightforward in principle to broaden our constraints to encompass all triggers produced by the pipelines. However, we refrain from doing so to keep our computational costs manageable. Our preliminary analyses of slightly longer triggers (8s, 16s, and 32s) yield results qualitatively similar to those presented below.

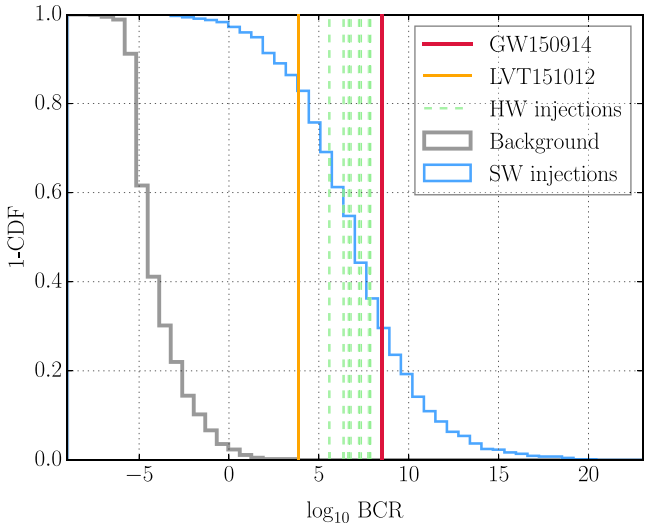

FIG. 1. BCR distributions. Histograms represent the survival function (1-CDF) from our selection of 983 aLIGO O1 background triggers (gray) and 648 simulated signals (blue). Vertical lines mark the BCRs of eight hardware injections (dashed green), LVT151012 (leftmost orange line), and GW150914 (thick red line). Background triggers were selected to be uniformly distributed in log-IFAR, and 98% yield $\log \mathrm{BCR} < 0$.

We compare the BCRs from our background selection to several foreground triggers. The foreground includes eight hardware injections, which were performed by physically actuating the test masses of the detectors to simulate signals similar to GW150914 [38]. We also analyze a set of 648 software injections: simulated signals inserted in O1 data, with arbitrary sky location and orientation, and with masses and spins that span our priors (in particular, the luminosity distance distribution is uniform in volume with a cutoff at 2.5 Gpc). On top of these artificial triggers, we also compute the BCR for GW150914 [1] and LVT151012 [3]. The freedom provided by the α and β parameters in Eq. (1) may be used to minimize the overlap between the simulated-signal and background distributions; the results below correspond to values of $\alpha = 10^{-6}$ and $\beta = 10^{-4}$, but may be adjusted in future analyses (see the Appendix).

V. RESULTS

Figure 1 shows the BCR distributions obtained for background triggers and software injections. The figure also displays the values obtained for GW150914, LVT151012, and hardware injections, all of which show much stronger evidence for being coherent CBC signals, rather than incoherent glitches (high BCR). We find a clear

[2]There are $\mathcal{O}(10^7)$ background triggers in O1. The run time on a single background trigger using the LALInference implementation of nested sampling is usually between 1 to 5 hours.

[3]For example, see Table IV in Appendix B of [3], where Bayes factors computed with and without ROQ can be compared (the values in that example are close, but not identical due to differences in waveform approximants).

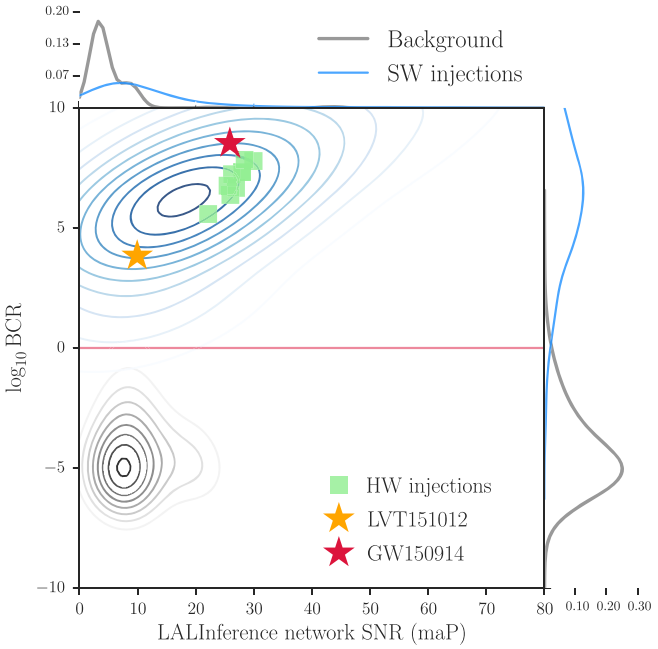

FIG. 2. BCR vs SNR distributions. Contours represent the normalized probability density of selected background triggers (gray) and simulated signals (blue) in log-BCR vs SNR space. The plot also shows eight hardware injections (green squares), LVT151012 (orange star), and GW150914 (red star). The curves shown on the right (top) result from a Gaussian kernel-density estimation of the one-dimensional distribution of log-BCRs (SNRs), obtained after integration over the x-axis (y-axis). Background triggers were selected to be uniformly distributed in log-IFAR, and 98% yield $\log \mathrm{BCR} < 0$ (threshold marked by a horizontal red line for convenience). The SNR on the x-axis is the coherent matched-filter signal-to-noise ratio of the template recovered with maximum *a posteriori* probability (maP) by our inference pipeline (LALINFERENCE).

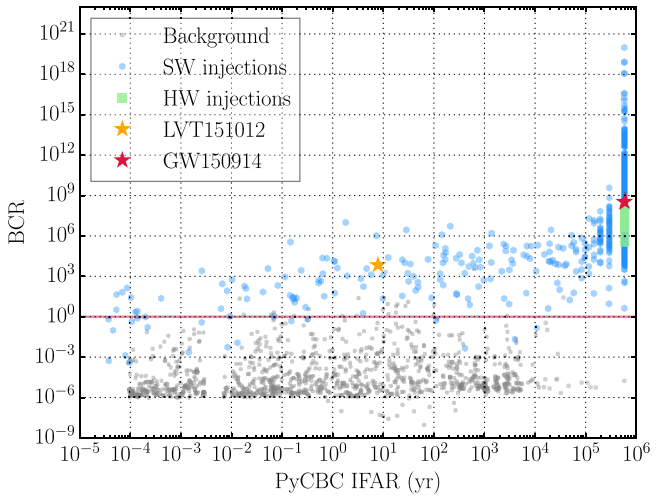

FIG. 3. BCR vs IFAR. BCR for the same data shown in Fig. 1, plotted vs the inverse false-alarm rate (IFAR) assigned to each event by PYCBC, one of the staple aLIGO search pipelines. There are six background triggers with BCR $\ll 10^{-9}$, which fall outside the range of this plot; no foreground triggers are excluded from this plot. High-significance events pile up on the right because their IFAR is a lower limit determined by the most significant trigger in the background. This plot suggests the BCR may be used to more easily reject incoherent glitches.

separation between injections and background events—suggesting that the BCR is good at distinguishing CBC signals from glitches. If we consider the intrinsic probabilistic meaning of the BCR, a value of $\log \mathrm{BCR} < 0$ indicates a preference for the instrumental-artifact hypothesis ($\mathcal{H}_\mathrm{I}$) over the coherent-signal one ($\mathcal{H}_\mathrm{S}$). As expected, the vast majority (98%) of background triggers fall bellow this mark, while the opposite is true for injections. GW150914 and LVT151012 yield $\log_{10} \mathrm{BCR}$ values of 8.5 and 3.8 respectively.

Figure 2 shows the same populations from Fig. 1, plotted also as a function of the network signal-to-noise ratio (SNR) recovered by our coherent Bayesian analysis. Figure 2 reveals that the BCR values of the signal population are correlated with SNR, which reflects the fact that we are better able to evaluate the coherence of signals that stand clearly above the noise floor. As a result, the separation between our signal and glitch populations improves with SNR. Because this population of background triggers was purposely selected to be uniform in log-IFAR, the gray contours in Fig. 2 should not be taken to be representative of the actual glitch distribution: this would include *vastly* more low-SNR triggers. In any case, BCR is largely independent of SNR for background triggers.

There are three software injections with SNR > 12, but BCR < 1. This is due to two characteristics that make the noise model preferable: (i) the ratio of SNRs in two detectors is greater than three, and (ii) the signal in at least one detector is too weak to be confidently discernible from Gaussian noise (SNR ~ 5.5). These rare circumstances are caused by source locations and orientations unfavorable to the detector network, and, as such, should be mitigated by adding more instruments.

Irrespective of its Bayesian interpretation, we may treat the BCR as a traditional detection statistic to obtain a frequentist estimate of the significance of any given foreground event based on the measured background (e.g., a p-value, or better, a likelihood ratio). Again, our background triggers were selected to represent common and rare events in equal numbers, so the distribution in Fig. 1 need not be the same as that of the entire background, and should not be used for this purpose. However, as shown in Fig. 3, we find that there is no evidence for strong correlation between BCR and the IFAR assigned by the detection pipeline. This suggests that the background BCR distribution shown in Fig. 1 is likely representative of the whole. Furthermore, Fig. 3 implies that the BCR may be used to more easily reject incoherent glitches, irrespective of IFAR,

and thus increase our detection confidence for marginal events like LVT151012.

VI. FUTURE IMPLEMENTATION

Given its ability to separate signals from glitches, the BCR may supplement existing search strategies and help increase their sensitivity, even with existing computational resources. The most straightforward way to achieve this would be to run existing CBC pipelines as usual, with an extra threshold on BCR (e.g., discarding any triggers with, say, BCR < 1). Our results suggest that this would be an efficient way of discarding the vast majority of instrumental artifacts, thereby increasing detection confidence of real signals [39].

Computational costs would currently preclude obtaining BCRs for *all* triggers (foreground and background) produced during a regular observation run, so this extra step would have to be reserved for the most significant ones, as determined by the main pipeline. However, processing all triggers *would* have the added advantage of potentially enabling the detection of weak GW events that would otherwise be missed (e.g., low-IFAR, but high-BCR, injections in Fig. 3). In the future, this would also enable us to move beyond a simple BCR veto, and instead use large numbers of simulated signals and background events to define empirical probability distributions over a space of multiple figures of merit (e.g., BCR and SNR, as in Fig. 2). This could be used to obtain likelihood ratios to categorize a trigger as signal or noise—which can be shown to be an optimal strategy for classification problems such as this, and have been used successfully by some existing searches [13,14,24]. Future improvements in ROQ methods, like their implementation on graphical processing units, will be vital in making this possible.

The values of the α and β weights in Eq. (1) have a strong effect on the shape of the distributions of Fig. 2, as discussed in Appendix. While here we have set them to values that yield a good separation between the signal and background populations, future studies may systematically optimize these parameters using a more comprehensive set of software injections and a large, representative set of background triggers. This may be achieved via any standard optimization scheme that attempts to minimize the overlap between the two populations. The values would, of course, be fixed before analyzing any foreground data.

VII. CONCLUSION

We have demonstrated that Bayesian models based on the coherence of GW triggers across detectors may successfully distinguish between real CBC signals and transient instrumental noise (Figs. 1 and 2). We introduced a specific figure of merit, the BCR, which responds to incoherent glitches in a way that is complementary to that of standard CBC pipelines (Fig. 3). Finally, we suggested a few avenues for incorporating this (or a similar) measure of coherence into existing GW search strategies, the simplest of which would take the form of a new veto for detection candidates. This could be implemented today to increase the number of gravitational waves confidently detected by LIGO and Virgo, without needing to further improve detector hardware.

Versions of the ranking statistic used by PyCBC in recent analyses have incorporated some measure of coherence [15], and it remains to be seen whether this introduces some correlation between BCR and IFAR in Fig. 3. Furthermore, while this study focused on detection candidates produced by the two aLIGO detectors during O1, we are currently investigating how the power of the BCR is affected by the addition of new detectors, like Virgo. Finally, although here we focused on short-duration (4s) triggers from high-mass binary-black-hole mergers, our preliminary results on slightly longer triggers (8s, 16s, and 32s) show qualitatively similar behavior.

ACKNOWLEDGMENTS

We thank Alan Weinstein, Alex Nitz, Carl-Johan Haster, Stefan Hild, Reed Essick, Ryan Lynch, Colm Talbot, Eric Thrane, John Veitch, and Thomas Dent for helpful comments. Rory Smith is supported by the Australian Research Council Centre of Excellence for Gravitational Wave Discovery (OzGrav), through Project No. CE170100004. The authors thank the LIGO Scientific Collaboration for access to the data and gratefully acknowledge the support of the United States National Science Foundation (NSF) for the construction and operation of the LIGO Laboratory and Advanced LIGO Grant No. PHY-0757058, as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, and the Max-Planck-Society (MPS) for support of the construction of Advanced LIGO. Additional support for Advanced LIGO was provided by the Australian Research Council. This manuscript has LIGO Document ID LIGO-P1700414.

APPENDIX: EFFECT OF BCR WEIGHTS

The weights (α, β) that go into the calculation of the BCR in Eq. (1) have a critical impact on the degree of separation between the signal and glitch populations. Here we elaborate on this point, and show how we improve upon previous work by explicitly taking advantage of the extra freedom afforded by these parameters.

From a Bayesian perspective, α and β encode our prior beliefs on the relative probabilities of each of the submodels that are compared in the computation of the BCR: α determines by what factor the coherent-signal hypothesis ($\mathcal{H}_S$) should be favored over the instrumental-feature hypothesis ($\mathcal{H}_I$),

$$\alpha \equiv \frac{P(\mathcal{H}_S)}{P(\mathcal{H}_I)}, \tag{A1}$$

while β gives the probability of the glitch hypothesis ($\mathcal{H}_G$) conditional on the assumption that there is an instrumental-feature to begin with,

$$\beta \equiv P(\mathcal{H}_{Gi}|\mathcal{H}_I) = 1 - P(\mathcal{H}_{Ni}|\mathcal{H}_I), \tag{A2}$$

for any dectector i, as discussed in Sec. III. The last equality in Eq. (A2) uses the fact that we *define* the instrumental-feature hypothesis as the logical union of the glitch and Gaussian noise ($\mathcal{H}_N$) subhypotheses, i.e., $\mathcal{H}_I \equiv \mathcal{H}_G \vee \mathcal{H}_N$, and that the latter are logically disjoint, i.e., $\mathcal{H}_G \wedge \mathcal{H}_N =$ False, so $P(\mathcal{H}_N|\mathcal{H}_G) = P(\mathcal{H}_G|\mathcal{H}_N) = 0$.

It follows from the probabilistic interpretation of these parameters that their allowed ranges are $0 < \alpha < \infty$ and $0 \leq \beta \leq 1$. All results presented in the main text were produced using the values

$$\text{Main text:}\ (\alpha = 10^{-6}, \beta = 10^{-4}). \tag{A3}$$

This specific choice was made to yield a good separation between the background and foreground populations, as reflected by Figs. 1 and 2. These values also result in an overall normalization such that BCR $= 1$ gives the point at which both hypotheses are equally likely given *our* trigger set (i.e., the horizontal red line in Fig. 2 roughly agrees with the intersection of the blue and gray curves on the right panel).

To see how α and β impact the separation between the background and foreground populations, consider as a proxy the distance between the mean BCRs for the two populations. In particular, define the quantity

$$\Delta_{b-f}\langle \log \text{BCR}\rangle \equiv \langle \log \text{BCR}^{(b)}\rangle - \langle \log \text{BCR}^{(f)}\rangle, \tag{A4}$$

where the angle brackets on the right denote averaging over triggers, and the superscripts (b) and (f) refer to background and foreground respectively. This number then gives a measure of the vertical distance between the centers of the distributions in Fig. 2. The effect of α and β on this quantity is shown in Fig. 4, where darker colors correspond to greater absolute mean distance. As expected from Eq. (1), the separation is a strong function of β, while it is largely independent of α. It can also be seen from Eq. (1) that α should merely impact the overall normalization of the BCR, shifting all values up or down.

By tuning β we may thus control the degree of bias introduced in the computation of the BCR. This can be used to correct for shortcomings in the definitions of the noise submodels themselves, so as to best distinguish foreground and background. The reason this is necessary in the first place is that not all glitches will conform strictly to the "worst-glitch" hypothesis as we have defined it via Eq. (3). For instance, the distribution of glitch morphologies and SNRs need not conform to the parameter priors assumed in the computation of Z^G; instead of tuning the parameter priors, one may correct for this effect via β (which is easier to implement).

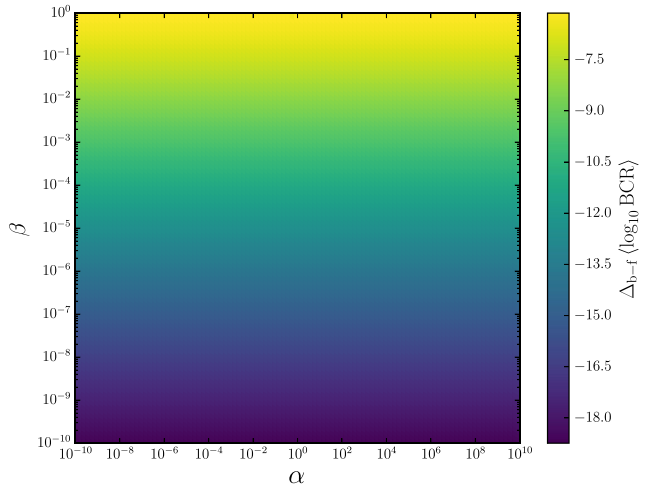

FIG. 4. Effect of weight on population separation. Color represents the difference in mean log BCR between background and foreground, $\Delta_{b-f}\langle \log \text{BCR}\rangle$ as defined in Eq. (A4). This is shown as a function of the BCR prior weights, α (x-axis) and β (y-axis), of Eq. (1). All values are negative because the foreground always has a larger mean, so darker colors correspond to greater distance between the population means.

Looking at Fig. 4, one may be tempted to substantially reduce β to maximize the distance between the distribution means. However, the quantity plotted in Fig. 4, Eq. (A4), is insensitive to the fact that the two distributions do not retain their shape when β is varied, and therefore is only useful as a proxy for population overlap when looking at small changes in the weights. In other words, Fig. 4 fails to convey the fact that there is a penalty in introducing too strong of a bias through β. This is related to the bias-variance trade-off, well known in statistical inference (see e.g., [40]). Let us explore how this trade-off is manifested throughout the range of valid values for β.

On one end, setting $\beta = 0$ comes at the price of throwing away all information about the incoherence of the trigger. As can be deduced from Eq. (1), in the limit of vanishing β the BCR is nothing but the usual signal vs Gaussian-noise odds (BSN),

$$\text{BCR}(\alpha = 1, \beta = 0) = Z^S/Z^N \equiv \text{BSN}, \tag{A5}$$

and the glitch model is totally ignored. For this choice of β, the BCR will just follow the usual dependence of BSN on SNR (see, e.g., [41]),

$$\log \text{BSN} \propto \text{SNR}^2, \tag{A6}$$

irrespective of whether the trigger is a glitch or a coherent signal, as shown in Fig. 5. Although the distance between the means of the two populations in this figure is large (as reflected also by Fig. 4, for $\beta \to 0$), this is only because, on

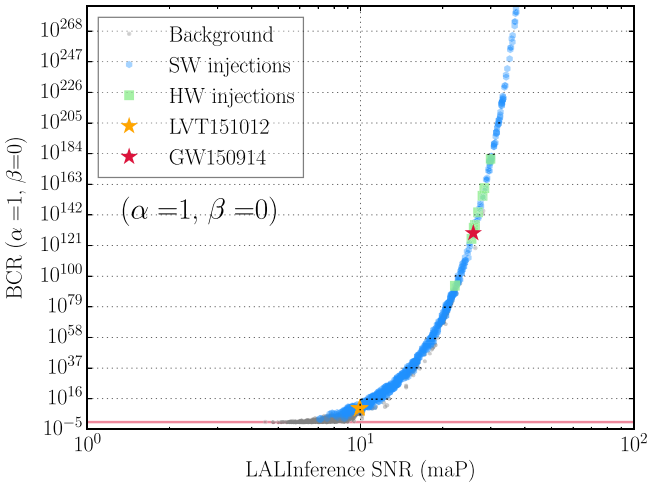

FIG. 5. BCR ($\alpha = 1, \beta = 0$) vs SNR. BCR vs SNR for the same data shown in Figs. 1–3, but analyzed with ($\alpha = 1$, $\beta = 0$). For this choice of weights, the BCR reduces to the Bayesian odds between signal and Gaussian noise, Eq. (A5), and scales with SNR according to Eq. (A6), for both background (gray circles) and foreground (blue hexagons). The SNR on the x-axis is the coherent matched-filter signal-to-noise ratio of the template recovered with maximum *a posteriori* probability (maP) by our inference pipeline (LALINFERENCE).

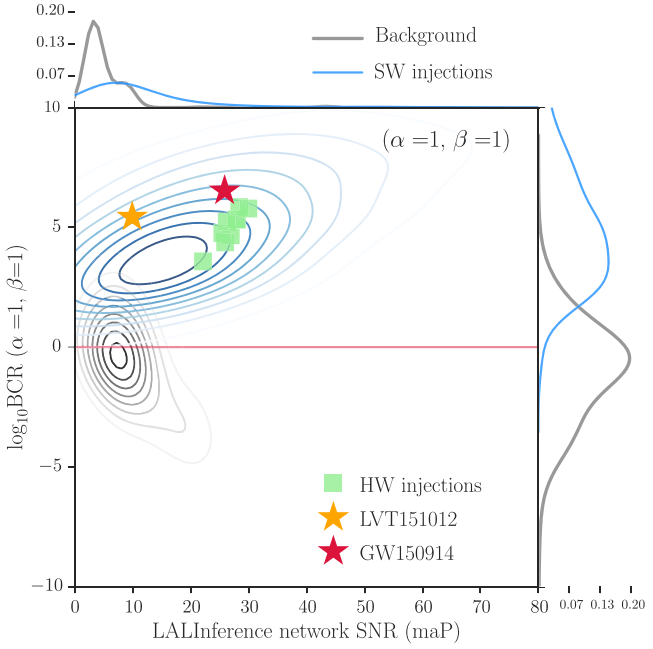

FIG. 6. BCR ($\alpha = 1, \beta = 1$) vs SNR distributions. This plot is completely analogous to Fig. 2, but with ($\alpha = 1$, $\beta = 1$) instead of ($\alpha = 10^{-6}$, $\beta = 10^{-4}$) [cf. Eq. (1)]. For this choice of weights, the BCR reduces to the BCI, Eq. (A7), resulting in greater overlap between the background (gray) and foreground (blue) distributions. For more details about this plot, refer to the caption of Fig. 2.

average, the background triggers in our set have lower SNR than the foreground.

On the other end, setting $\beta = 1$ is equivalent to ignoring the possibility that the trigger was produced by Gaussian noise. In that case, the BCR reduces to the evidence ratio between the coherent-signal and incoherent-glitch hypotheses, a quantity often called "BCI" by gravitational-wave data analysts (assuming $\alpha = 1$):

$$\mathrm{BCR}(\alpha = 1, \beta = 1) = Z^S/Z^G \equiv \mathrm{BCI}. \qquad \text{(A7)}$$

The use of this quantity for glitch-discrimination purposes in CBC searches was proposed in [11]. However, we find that it does not produce a sufficient separation between the background and foreground populations, except for loud triggers. For example, while ($\alpha = 10^{-6}$, $\beta = 10^{-4}$) yields Fig. 2, ($\alpha = 1, \beta = 1$) yields Fig. 6. From this plot, it is easy to see that the BCI is good at distinguishing *loud* incoherent glitches from *loud* coherent signals, but is inconclusive for weak triggers.

We can check that changing β indeed affects primarily *weak* glitches by comparing Fig. 7 to Fig. 3, BCR vs IFAR plots which were produced with $\beta = 1$ and $\beta = 10^{-4}$, respectively. The change in β from Fig. 7 to Fig. 3 causes low-IFAR (low-SNR) glitches to yield significantly lower BCRs, while high-IFAR (high-SNR) triggers are largely unaffected. Importantly, low-IFAR (low-SNR) signals are also down-ranked after the change, but to a lesser degree on average. Hence the separation in BCR improves, as can be seen by comparing the right panels of Figs. 6 and 2.

To further quantify the effect of β, we can also look at the fractional change in log BCR when going from ($\alpha = 1$, $\beta = 1$) to ($\alpha = 10^{-6}$, $\beta = 10^{-4}$),

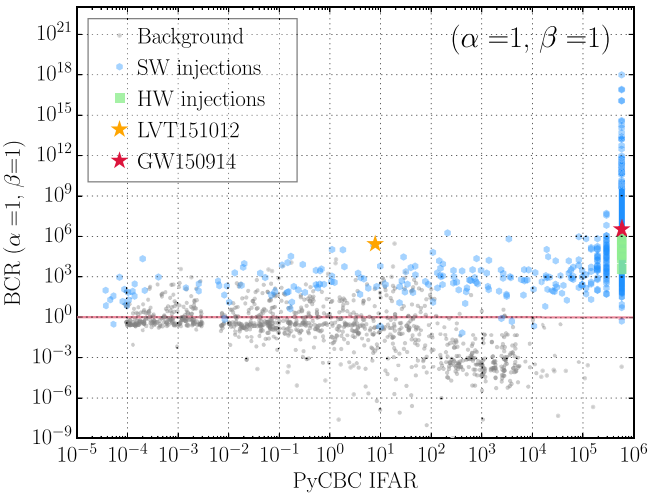

FIG. 7. BCR ($\alpha = 1, \beta = 1$) vs IFAR. This plot is completely analogous to Fig. 3, but with ($\alpha = 1, \beta = 1$) instead of ($\alpha = 10^{-6}$, $\beta = 10^{-4}$) [cf. Eq. (1)]. For this choice of weights, the BCR reduces to the BCI, Eq. (A7), resulting in greater overlap between the background (gray) and foreground (blue) distributions. For more details about this plot, refer to the caption of Fig. 3.

$$\frac{\Delta(\log \mathrm{BCR})}{|\log \mathrm{BCI}|} \equiv \frac{\log \mathrm{BCR}(10^{-6}, 10^{-4}) - \log \mathrm{BCI}}{|\log \mathrm{BCI}|}, \quad \text{(A8)}$$

where vertical bars mark absolute values, and the BCI is defined by Eq. (A7). This quantity is histogrammed in Fig. 8 for the triggers in our set. The fact that the change in β affects weak glitches more significantly than strong ones is reflected in the bimodality of the gray distribution: the left (right) peak corresponds to triggers below (above) an effective threshold of SNR ~ 9. On the other hand, the blue distribution in Fig. 8 shows that most (although not all) signals are largely unaffected by the change in β, with a mean increase in BCR but long tails extending mainly to the left. This large variance is due mostly to the weaker signals for which the BCR decreased due to the change in β.

By tuning the weights, we may attempt to find a sweet spot in which the bias introduced is just enough to separate weak glitches from weak signals, without confounding loud glitches with loud signals. The choice of Eq. (A3) was found to be close to this ideal, and achieves this by separating the weak glitches in our set from the weak signals to an extent, largely without altering loud triggers (Figs. 1–3).

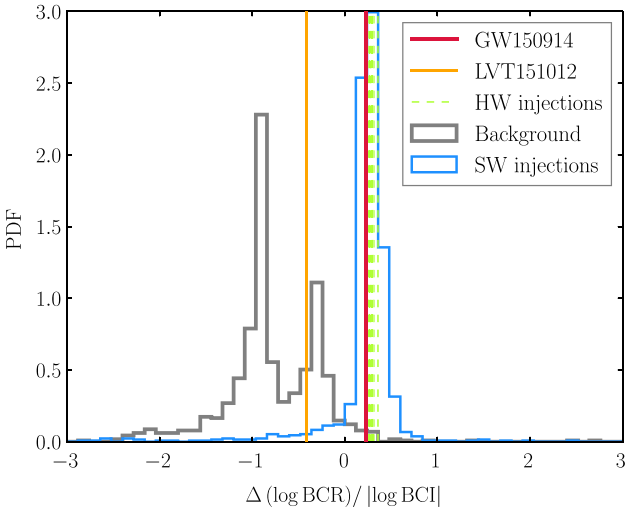

FIG. 8. Effect of weights on log BCR. Histogram of the fractional change in log BCR when going from ($\alpha = 1$, $\beta = 1$) to ($\alpha = 10^{-6}$, $\beta = 10^{-4}$), Eq. (A8). This plot summarizes the differences between the BCRs shown in Figs. 2 and 3 and those in Figs. 6 and 7.

[1] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **116**, 061102 (2016).
[2] B. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **116**, 241103 (2016).
[3] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. X **6**, 041015 (2016).
[4] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **118**, 221101 (2017).
[5] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **119**, 141101 (2017).
[6] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **119**, 161101 (2017).
[7] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **120**, 091101 (2018).
[8] J. Aasi *et al.* (LIGO Scientific Collaboration), Classical Quantum Gravity **32**, 115012 (2015).
[9] F. Acernese *et al.* (Virgo Collaboration), Classical Quantum Gravity **32**, 024001 (2015).
[10] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Classical Quantum Gravity **33**, 134001 (2016).
[11] J. Veitch and A. Vecchio, Phys. Rev. D **81**, 062003 (2010).
[12] K. Cannon *et al.*, Astrophys. J. **748**, 136 (2012).
[13] K. Cannon, C. Hanna, and D. Keppel, Phys. Rev. D **88**, 024025 (2013).
[14] C. Messick *et al.*, Phys. Rev. D **95**, 042001 (2017).
[15] A. H. Nitz, T. Dent, T. Dal Canton, S. Fairhurst, and D. A. Brown, Astrophys. J. **849**, 118 (2017).
[16] S. A. Usman *et al.*, Classical Quantum Gravity **33**, 215004 (2016).
[17] T. Dal Canton *et al.*, Phys. Rev. D **90**, 082004 (2014).
[18] M. Isi, M. Pitkin, and A. J. Weinstein, Phys. Rev. D **96**, 042001 (2017).
[19] J. Skilling, Bayesian Anal. **1**, 833 (2006).
[20] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).
[21] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, Phys. Rev. D **85**, 082003 (2012).
[22] D. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial*, 2nd ed. (Oxford University Press, New York, 2006).
[23] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, 2003).
[24] R. Lynch, S. Vitale, R. Essick, E. Katsavounidis, and F. Robinet, Phys. Rev. D **95**, 104046 (2017).
[25] N. J. Cornish and T. B. Littenberg, Classical Quantum Gravity **32**, 135012 (2015).
[26] J. Powell, M. Szczepanczyk, and I. S. Heng, Phys. Rev. D **96**, 123013 (2017).
[27] D. Keitel, R. Prix, M. A. Papa, P. Leaci, and M. Siddiqi, Phys. Rev. D **89**, 064023 (2014).
[28] M. Pitkin, C. Gill, J. Veitch, E. Macdonald, and G. Woan, J. Phys. Conf. Ser. **363**, 012041 (2012).
[29] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Astrophys. J. **839**, 12 (2017).
[30] A. Nitz, I. Harry, D. Brown *et al.*, ligo-cbc/pycbc: Latest release, 2017, DOI: 10.5281/zenodo.1313589.
[31] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **116**, 241102 (2016).

[32] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, Phys. Rev. D **94**, 044031 (2016).
[33] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016).
[34] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).
[35] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).
[36] T. B. Littenberg and N. J. Cornish, Phys. Rev. D **91**, 084034 (2015).
[37] N. J. Cornish and T. B. Littenberg, Classical Quantum Gravity **32**, 135012 (2015).
[38] C. Biwer, D. Barker, J. C. Batch *et al.*, Phys. Rev. D **95**, 062002 (2017).
[39] J. B. Kanner, T. B. Littenberg, N. Cornish, M. Millhouse, E. Xhakaj, F. Salemi, M. Drago, G. Vedovato, and S. Klimenko, Phys. Rev. D **93**, 022002 (2016).
[40] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer, New York, New York, NY, 2009).
[41] J. D. E. Creighton and W. G. Anderson, *Gravitational-Wave Physics and Astronomy* (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2011), p. 188.